Biophysical *Journal*
Computational Tool

Biophysical Society

ANCA: Anharmonic Conformational Analysis of Biomolecular Simulations

Akash Parvatikar,[1] Gabriel S. Vacaliuc,[2] Arvind Ramanathan,[2] and S. Chakra Chennubhotla[1,*]
[1]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania and [2]Computational Science and Engineering Division, Health Data Sciences Institute, Oak Ridge National Laboratory, Oak Ridge, Tennessee

ABSTRACT Anharmonicity in time-dependent conformational fluctuations is noted to be a key feature of functional dynamics of biomolecules. Although anharmonic events are rare, long-timescale (μs–ms and beyond) simulations facilitate probing of such events. We have previously developed quasi-anharmonic analysis to resolve higher-order spatial correlations and characterize anharmonicity in biomolecular simulations. In this article, we have extended this toolbox to resolve higher-order temporal correlations and built a scalable Python package called anharmonic conformational analysis (ANCA). ANCA has modules to: 1) measure anharmonicity in the form of higher-order statistics and its variation as a function of time, 2) output a storyboard representation of the simulations to identify key anharmonic conformational events, and 3) identify putative anharmonic conformational substates and visualization of transitions between these substates.

INTRODUCTION

Traditional analysis tools for biomolecular simulations have focused on second-order statistics (1–3). Anharmonicity in time-dependent conformational fluctuations is noted to be a key feature of functional dynamics of biomolecules (4–6). Although anharmonic events are rare, long-timescale (μs–ms and beyond) simulations facilitate probing their behavior. However, automated analyses and visualization of anharmonic events from these long-timescale simulations are proving to be a significant bottleneck.

We have addressed this challenge previously by proposing anharmonicity as an organizing principle for conformational landscapes of proteins and other biomolecules (7). In particular, we have built a quasi-anharmonic analysis toolbox to resolve higher-order *spatial* correlations (8–11). In this work, we have extended this toolbox to resolve higher-order *temporal* correlations from long-timescale simulations and built a scalable Python package, anharmonic conformational analysis (ANCA). ANCA has modules to: 1) measure anharmonicity in the form of higher-order statistics and its variation as a function of time, 2) output a storyboard representation of the simulations to identify key anharmonic conformational events, and 3) identify putative anharmonic conformational substates and visualization of transitions between these substates.

Submitted October 19, 2017, and accepted for publication March 8, 2018.

*Correspondence: chakracs@pitt.edu

Editor: Nathan Baker.
https://doi.org/10.1016/j.bpj.2018.03.021

Description and functionality

Inputs to ANCA

ANCA can process trajectories in many formats commonly used by the biophysics community, including Protein Data Bank, CHARMM DCD files, AMBER coordinates, and Gromacs xtc files. ANCA uses MDAnalysis (12,13) and mdtraj (14) to capture and process coordinate (or other feature) information from molecular dynamics (MD) trajectory files. Further, the user can specify which features to select and process using an extensive set of coordinate and feature selection commands within the two packages. Using Python's inbuilt capabilities to process memory-mapped arrays, we can process large trajectories up to several terabytes. We demonstrate ANCA in analyzing a publicly available millisecond-long trajectory data of the protein bovine pancreatic trypsin inhibitor (BPTI) (15).

Conformational events storyboard

Using κ to quantify anharmonicity in positional/angular deviations within MD simulations. To complement insights from harmonic measures of conformational changes such as the root mean-squared deviation, we have used higher-order anharmonic measures, namely kurtosis (κ) (8). κ is

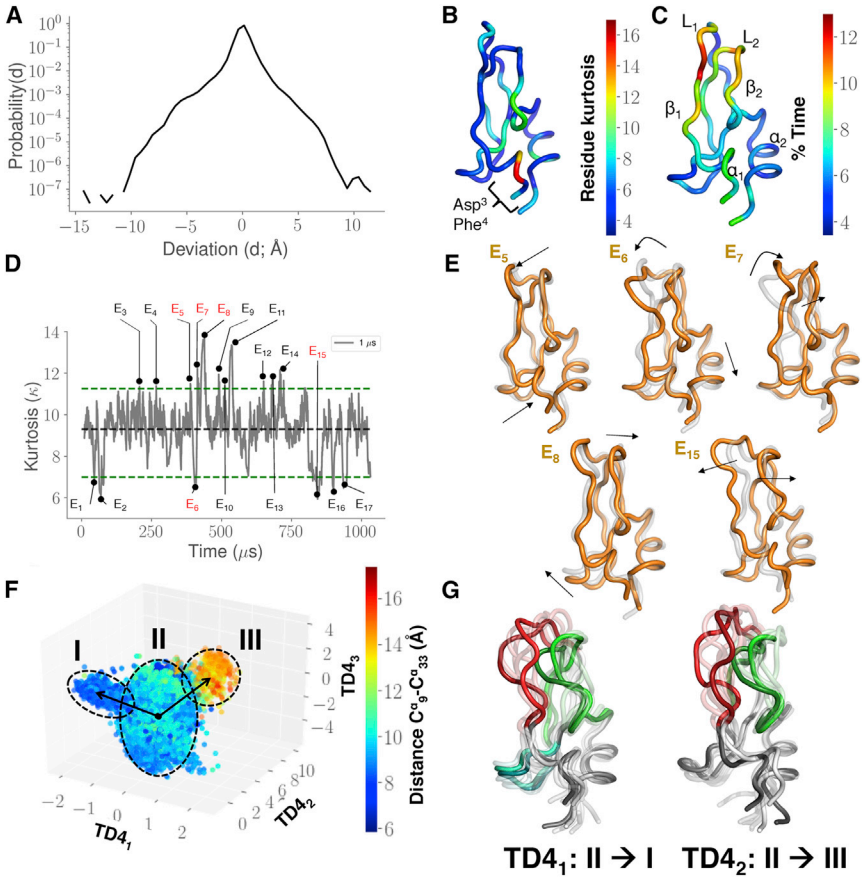

FIGURE 1 ANCA analysis of a millisecond-long simulation of BPTI. (*A*) The positional deviations of C^{α} atoms are anharmonic and long-tailed (κ = 15.94; *z*-score = 3778.44 and *p*-value = 0.00). (*B*) Residues are colored by individual kurtosis (κ) values. Two residues—Asp^3-Phe^4—show the largest κ values while sampling anharmonic motions infrequently, as shown in (*C*). Fig. S1 provides additional details on tracking the conformational events for these two residues. The anharmonic fluctuations can lead to significant conformational changes, as shown in (*D*) and (*E*). (*D*) The time evolution of κ values seen through an exponential sliding window of 1-μs half-life. Using a threshold of four SDs (*green dotted lines*) above and below the mean κ (*black dotted line*), we identify a total of 17 conformational events, labeled $\mathcal{E}_1 - \mathcal{E}_{17}$. (*E*) We show five select events, $\mathcal{E}_5$, $\mathcal{E}_6$, $\mathcal{E}_7$, $\mathcal{E}_8$, and $\mathcal{E}_{15}$ as ensembles, with the gray cartoon representing the previous event and the orange cartoon representing the current event. Arrows are used to highlight the opening/closing of the flap regions of BPTI between events. (*F*) A multidimensional description of the simulation data using the top three time-delayed anharmonic modes is given. Each conformation, represented by a dot, is colored by the distance between the centers-of-mass of the flap regions (L1 and L2 in (*C*)). Three putative conformational substates are demarcated by dotted ellipses depicting the closed (*I*) and open (*III*) states that pass through an intermediate state (*II*), as seen by the colored distance distribution. The arrows indicate how to reach the closed and open states by walking along anharmonic modes $TD4_1$ and $TD4_2$ from the intermediate state. (*G*) These motions are shown in an ensemble form, with L_1 (*red*), L_2 (*green*), $\beta_1 - \beta_2$ (*cyan*), and the rest of the protein (*gray*) depicted in light to dark colors, denoting start-to-end trajectory evolution.

calculated from either the Cartesian coordinates or dihedral angle selections specified by the user. For a unimodal Gaussian distribution with zero mean and unit variance, $\kappa = 3$; a value of $\kappa > 3$ indicates a super-Gaussian distribution that is more peaked and heavier-tailed than the baseline Gaussian. Conversely, a distribution that is less peaked than the baseline Gaussian has kurtosis $\kappa < 3$. The statistical significance of κ is assessed through the kurtosis test, which rejects the hypothesis of normality when the p-value <0.05. Fig. 1 *A* shows the histogram of positional deviations of C^{α} atoms in the BPTI simulation. Using κ, we quantify which parts of the protein exhibit anharmonic motions (Fig. 1 *B*) and for how long (Fig. 1 *C*). In the case of BPTI, we can observe that a majority of the C^{α} atoms spend at least 5% of their time exhibiting anharmonic motions. However, helix two is mostly harmonic because of the strong hydrophobic interactions and Cys-disulfide bonds.

We analyzed the variation of κ over the length of the trajectory at each C^{α} coordinate (x, y, z) using an exponential window with a half-life of 1 μs (11). Almost all of the individual residues exhibit some degree of anharmonicity (Table S1), whereas κ is more pronounced along individual coordinate directions (Table S2). These conformational changes constitute events within the trajectory that may be of interest to the user for further analysis.

Kurtosis-based event detection. Using κ, the user can identify conformational events that occur at distinct timescales (by changing the half-life of the exponential window) and organize a conformational storyboard for the entire simulation(s). Fig. 1 *D* shows the variation of kurtosis over time using an exponential window with a half-life of 1 μs; the filtering procedure is described in detail in (11). Using a user-defined threshold (*green line* in Fig. 1 *D*), a total of 17 conformational events are detected (labeled $\mathcal{E}_1$–$\mathcal{E}_{17}$). Select events from this are organized as a storyboard in Fig. 1 *E*. These events summarize the time points at which the BPTI loops L_1 and L_2 open/close. The storyboard provides a means to quickly summarize large MD trajectories while allowing the user to visually interact with events of interest and simultaneously track other quantities of interest (e.g., root mean-squared deviation, R_g, etc.) over the course

of long simulations (data not shown). In addition to using κ, conformational events can be detected with information theoretic measures such as mutual information (16); however, these techniques can be computationally expensive. Trajectory segments from the storyboard can be further analyzed to identify putative conformational substates, as discussed below. We also provide the ability to construct storyboards for individual residues (see Fig. S1 for an illustration).

Characterizing anharmonic modes of motion in the conformational landscape

ANCA provides four core modules for analyzing MD trajectories. These modules take as input X either Cartesian coordinates of dimensions $3N \times t$, where $3N$ represents the three-dimensional (x, y, z) coordinates of the individual atoms selected for analysis, or cosine/sine transformed dihedral angles, namely (ϕ, ψ, χ) resulting in a $D \times t$, where D represents the total number of transformed dihedral angle selections. In both cases, t represents the total number of conformations from the simulations.

The SD2 module removes dominant second-order spatial correlations by computing a spatial covariance matrix and performing principal component analysis. In addition to the simulation data, SD2 requires as input m the subspace dimensionality. m can be adjusted by examining the inflection points in the cumulative variance plots that this module returns. SD2 diagonalizes the covariance matrix and returns the eigenvalues S (size $m \times 1$), eigenvectors B ($3N$ or $D \times m$), and the projection matrix $Y = B^T X$ ($m \times t$). The top three modes from the SD2 module for the BPTI simulations are shown in Figs. S2 *A* and S3.

The SD4 module (previously quasi-anharmonic analysis (8)) attempts to resolve the intrinsic nonorthogonal spatial dependencies in atomistic fluctuations. The second-order projections, Y, from SD2 are used to build a fourth-order spatially correlated cumulant tensor. SD4 approximately diagonalizes this tensor to return an anharmonic mode matrix W ($3N$ or $D \times m$). The default ordering of the ANCA modes is based on the kurtosis of the projected coordinates; however, this ordering may not always correspond to a biophysically relevant reaction coordinate (11). This can be attributed to the fact that ANCA pursues rare conformational events, and if the projected coordinates correlate with such rare events, then ANCA can indeed provide biophysically meaningful projections.

To build associations between the SD4 modes and biophysically meaningful reaction coordinates, the user can upload physical observables such as radius of gyration (R_g), pairwise distances between specific atoms/groups of atoms, or overall energy values (potential + kinetic) from the simulations and simultaneously visualize how the physical observables map onto each of the SD4 modes (8) or use other techniques to identify reaction coordinates (17). For the BPTI simulations, the top three modes from the SD4 module are shown in Figs. S2 *B* and S4. We used the distance between residues Pro^9 and Phe^{33} to map the conformational fluctuations involved in opening/closing of the BPTI flaps. Indeed, the motions captured by $SD4_3$ correspond to an increase in the distance between the flap regions of BPTI.

The TD2 module removes dominant second-order temporal correlations by computing a time-delayed covariance matrix and performing principal component analysis. The inputs to this module are similar to the SD2 module, with one additional user-specified parameter, τ, that denotes the lag time over which the temporal correlations are to be resolved. The outputs of this module include Z, a matrix obtained by projecting the simulation data on the dominant time-delayed eigenvectors and the corresponding eigenvalues. The top three modes from the SD4 module for the BPTI simulations are shown in Figs. S2 *C* and S5.

The TD4 module constructs a time-delayed fourth-order kurtosis tensor, which is then approximately diagonalized to obtain anharmonic modes of motions once the second-order spatial and temporal correlations are resolved (18). The TD4 module is the temporal analog of the spatial SD4 module. The input parameters to this module includes the matrix Z (from the TD2 module), a user-specified subspace value m denoting the number of desired anharmonic modes of motion, the lag time τ, and the matrix V. The outputs from the module include the separating matrix W.

For BPTI, the projections from the three principal TD4 modes ($TD4_1$–$TD4_3$) depicted in Fig. 1 *F* describe essential motions of the flap regions along two distinct directions. To quantify these motions, we use a reaction coordinate based on the distances between residues Pro^9 and Phe^{33}. To understand these motions further, we depict the conformational transitions in BPTI (Fig. 1 *G*); in each case, the flaps open/close, albeit in distinct directions and in some cases even capturing rare transitions involved in exchange of the flaps (see Supporting Material). The ANCA modes enable us to quantitatively understand the extent to which the relative motions between the flaps expose opening/closing of this region. The projections of the simulation data as well as the description of the principal modes from SD2, SD4, and TD2 for BPTI are provided in Figs. S3–S5.

Visualization

We provide the user with example iPython notebooks to visualize the results from the analyses over a web browser (Fig. 1, *A*, *D*, and *F*). To visualize structural data obtained from ANCA, we provide scripts for generating anharmonic modes using PyMOL or Visual Molecular Dynamics (Fig. 1, *B*, *C*, *E*, and *G*). Individual regions in the protein can be colored using the output PyMOL files. ANCA is available as an open-source Python package under the BSD 3-Clause License. Python tutorial notebooks, documentation and examples are available for download from http://csb.pitt.edu/anca.

Conclusion

Several applications support analyses of MD trajectories based on second-order statistics, including MDAnalysis (12,13) and mdtraj (14). To complement these tools, we have developed ANCA as a package for analyzing higher-order anharmonic motion signatures from MD simulations. ANCA provides a biophysically meaningful organizational framework for long-timescale biomolecular simulations and can be integrated with other software such as PyEMMA (19) to build Markov models of MD simulations.

SUPPORTING MATERIAL

Five figures and two tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)30388-6.

AUTHOR CONTRIBUTIONS

A.P. and G.S.V. implemented the analytical tools in Python. A.R. designed the software architecture and supervised its implementation. S.C.C. built the algorithmic framework for anharmonic analysis of conformational ensembles and conceptualized a conformational analysis toolkit around anharmonicity. All the authors wrote and reviewed the article.

ACKNOWLEDGMENTS

The authors thank the D. E. Shaw Research group for providing access to the MD simulations of BPTI. The work of S.C.C. and A.P. was supported by National Institutes of Health-National Institute of General Medical Sciences grant GM105978.

REFERENCES

1. Amadei, A., A. B. Linssen, and H. J. Berendsen. 1993. Essential dynamics of proteins. *Proteins.* 17:412–425.
2. Lange, O. F., and H. Grubmüller. 2006. Can principal components yield a dimension reduced description of protein dynamics on long time scales? *J. Phys. Chem. B.* 110:22842–22852.
3. Altis, A., P. H. Nguyen, …, G. Stock. 2007. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* 126:244111.
4. Mao, B., M. R. Pear, …, S. H. Northrup. 1982. Molecular dynamics of ferrocytochrome c: anharmonicity of atomic displacements. *Biopolymers.* 21:1979–1989.
5. Ichiye, T., and M. Karplus. 1987. Anisotropy and anharmonicity of atomic fluctuations in proteins: analysis of a molecular dynamics simulation. *Proteins.* 2:236–259.
6. Ichiye, T., and M. Karplus. 1988. Anisotropy and anharmonicity of atomic fluctuations in proteins: implications for X-ray analysis. *Biochemistry.* 27:3487–3497.
7. Ramanathan, A., A. Savol, …, P. K. Agarwal. 2014. Protein conformational populations and functionally relevant substates. *Acc. Chem. Res.* 47:149–156.
8. Ramanathan, A., A. J. Savol, …, C. S. Chennubhotla. 2011. Discovering conformational sub-states relevant to protein function. *PLoS One.* 6:e15827.
9. Savol, A. J., V. M. Burger, …, C. S. Chennubhotla. 2011. QAARM: quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin. *Bioinformatics.* 27:i52–i60.
10. Burger, V. M., A. Ramanathan, …, C. S. Chennubhotla. 2012. Quasi-anharmonic analysis reveals intermediate states in the nuclear co-activator receptor binding domain ensemble. In Proceedings of the Pacific Symposium on Biocomputing, R. B. Altman et al., eds. (Pacific Symposium on Biocomputing), pp. 70–81.
11. Ramanathan, A., A. J. Savol, …, C. S. Chennubhotla. 2012. Event detection and sub-state discovery from biomolecular simulations using higher-order statistics: application to enzyme adenylate kinase. *Proteins.* 80:2536–2551.
12. Michaud-Agrawal, N., E. J. Denning, …, O. Beckstein. 2011. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* 32:2319–2327.
13. Gowers, R., M. Linke, …, O. Beckstein. 2016. MDAnalysis: a Python package for rapid analysis of molecular dynamics simulations. In Proceedings of the 15th Python in Science Conference, S. Benthall and S. Rostrup, eds. (SciPy), pp. 98–105.
14. McGibbon, R. T., K. A. Beauchamp, …, V. S. Pande. 2015. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* 109:1528–1532.
15. Shaw, D. E., P. Maragakis, …, W. Wriggers. 2010. Atomic-level characterization of the structural dynamics of proteins. *Science.* 330:341–346.
16. McClendon, C. L., G. Friedland, …, M. P. Jacobson. 2009. Quantifying correlations between allosteric sites in thermodynamic ensembles. *J. Chem. Theory Comput.* 5:2486–2502.
17. Best, R. B., and G. Hummer. 2005. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA.* 102:6732–6737.
18. Georgiev, P., and A. Cichocki. 2003. Robust independent component analysis via time-delayed cumulant functions. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* 86:573–579.
19. Scherer, M. K., B. Trendelkamp-Schroer, …, F. Noé. 2015. PyEMMA 2: a software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* 11:5525–5542.